Víctor H. Escobar-Jeria

Universidad Tecnológica Metropolitana

vescobar@utem.cl

María J. Martín-Bautista

Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada 18071 Granada, España

mbautis@ decsai.ugr.es

María A. Vila

Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada 18071 Granada, España

vila@ decsai.ugr.es

PERFIL DE USUARIO Y LÓGICA DIFUSA: MODELO DE REPRESENTACIÓN DE PERFIL DE USUARIO EN XML

RESUMEN

Existen factores importantes que han llevado a un aumento de la información, como la acumulación rápida de datos, el desarrollo de sistemas gestores de base de datos más poderosos y el constante desarrollo tecnológico en el que Internet y las bases de datos dinámicas, entre otras, pasan a ser las principales fuentes de extracción de información. En el campo del KDD, la lógica difusa puede ayudar en los problemas tales como la manipulación de asuntos relacionados con la comprensibilidad de patrones, datos ruidosos e incompletos y la interacción humana. Concretamente, este documento se centra en las aplicaciones de lógica difusa en la Minería Web que permitan mejorar y optimizar diferentes procesos de la Web.

Como en la Minería Web de Uso existen diferentes elementos que tienen cierto grado de imprecisión o de incertidumbre, la lógica difusa permite manipular esos elementos para poder representar de mejor forma la realidad. Un ejemplo significativo son los perfiles de usuario. El tratamiento de los perfiles de usuario contempla el análisis de la información, que es registrada a partir de diversas fuentes, con el fin de obtener información sobre las preferencias y características de los usuarios. Muchas de estas preferencias o intereses tienen cierto grado de incertidumbre, por lo que la lógica difusa va a permitir modelar y manejar dicha información de una manera flexible.

Palabras clave: **Minería Web, Minería Web de Uso, Lógica Difusa, Personalización, Perfil de usuario, XML.**

INTRODUCCIÓN

En la actualidad, la información disponible en Internet cumple un rol muy importante en nuestra vida cotidiana, desde comprar por la Web o revisar nuestros correos electrónicos, hasta leer un simple periódico digital. Esto indica que con el tiempo ha sido más fácil obtener información, pero, a la vez, existe un gran desafío para las empresas y las organizaciones que tiene que ver con generar información de calidad, oportuna e interesante para los usuarios que navegan por la Web. Para manejar esta gran cantidad de información surge un área de estudios llamada

Minería Web. Esta área no es sólo una aplicación de técnicas de Minería de Datos para los datos de la Web, sino una adaptación de procesos más adecuados para las características de este tipo dato, principalmente por la falta de estructura.

Cuando los usuarios navegan por el Web van dejando huellas, estas huellas permiten de alguna manera identificar a un usuario o un grupo de usuarios con similares y distintas características. La gran mayoría de la información acerca de la actividad que realizan los usuarios en la Web es almacenada en los archivos log.

Uno de los objetivos de la Minería Web de Uso es minar los archivos log del servidor web para encontrar relaciones entre usuarios acerca de los aspectos de navegación. La mayor parte de estos aspectos son conceptos intrínsecamente imprecisos o inciertos. La edad del usuario, el nivel de educación, el lenguaje, qué tan paciente es el usuario en su navegación, son los conceptos fáciles de manejar a través de técnicas como la Lógica Difusa [23].

Además, la Lógica difusa es también utilizada en aplicaciones de Minería para mejorar las capacidades de alguna de las técnicas y algoritmos [2]; dos de las más extendidas son las Reglas de Asociación Difusas [11] y el Clustering Difuso [3].

Este artículo esta estructurado de la siguiente manera: en el apartado dos, explicaremos brevemente los tres tipos de Minería Web: Minería Web del Contenido, Minería Web de la Estructura y Minería Web del Uso, enfocándonos principalmente en la Minería Web en el apartado tres. Para este propósito hemos estudiado el proceso de personalización en la Minería Web a través de técnicas de minería, caracterización y clustering de sesiones de usuarios con el fin de identificar a grupos sociales de usuarios. En el apartado cuatro propondremos el modelo de representación de perfiles de usuarios y veremos algunos resultados prácticos obtenidos. Por último, en el apartado cinco veremos las conclusiones y futuros trabajos.

MINERÍA WEB

Actualmente, vivimos en un mundo tan involucrado con el Internet, que con un simple click podemos saber lo que sucede en otras partes del mundo en unos pocos segundos. Sin embargo, la situación se altera cuando queremos obtener información útil y precisa. El volumen grande de información en la Web, la diversidad de lenguajes, la calidad de información o la distribución de información en diferentes plataformas son algunas de las inconveniencias para procesar estos datos.

Por medio de las técnicas de Minería, nueva información puede ser descubierta y extraída automáticamente de documentos y servicios web [13]. Sin embargo, las técnicas de Minería tradicional no siempre son tan fáciles de aplicar en la estructura de la web. Las particularidades de esas fuentes suponen la falta de estructura y diversidad de representación, lo que determina que la etapa del preprocesamiento sea uno de los más importantes en el proceso de Minería [10], [15].

En [5] se distinguen tres formas para entender a la Minería Web, desde el punto de vista del contenido, estructura y uso. *La Minería Web del Contenido* es el descubrimiento automático de patrones de contenidos en documentos web [18], [6]; la *Minería Web de la Estructura* consiste en estudiar los enlaces de las estructuras entre o intra documentos para descubrir patrones útiles [9], [6]; y finalmente, la *Minería Web del Uso* está definido como el proceso de descubrimiento automático de patrones de accesos o uso de servicios web, basados en el comportamiento de los usuarios cuando interactúan en la Web [20].

MINERÍA WEB DEL USO

En el comercio electrónico es necesario aplicar métodos o técnicas inteligentes que permitan almacenar y analizar la información que los clientes o potenciales clientes van dejando registrada en su navegación por la Web. Por esta razón, el comportamiento del usuario y los objetivos del usuario son elementos importantes que se deben obtener. El conocimiento acerca del usuario es usado no sólo para caracterizar al usuario, sino también para descubrir tendencias genéricas para los propósitos de marketing y para la mejora del sitio Web. Estos datos acerca de los usuarios son coleccionados de la actividad del usuario en el web por medio de huellas digitales (IP, factors, cookies, etc.) que los servidores almacenan automáticamente en los accesos. Esta actividad es principalmente almacenada en los archivos log del servidor web, y es usualmente completada con otras fuentes informativas que provienen del servidor proxy, de la máquina del usuario, entre otros.

Para la explotación de diferentes tipos de datos, pueden ser usadas diferentes técnicas de descubrimiento de patrones. Estos patrones de navegación pueden evidenciar preferencias del usuario, el comportamiento del cliente en línea y futuras direcciones de mejoras del sitio Web.

La información que puede ser obtenida de estos procesos de Minería nos permitirá construir perfiles de usuario prototípicos y llevar a cabo un proceso de personalización. Obteniendo los perfiles del usuario, varias decisiones pueden ser tomadas desde el punto de vista de los contenidos y diseño para personalizar el sitio en función del usuario.

Personalización y perfil de usuario

La personalización de la Web puede ser vista desde dos puntos de vista: La empresa y el usuario. El punto de vista de la empresa está relacionado con el marketing e identificación de clases demográficas. El punto de vista del usuario está relacionado con la obtención de información y las recomendaciones de navegación. Parte de esta información puede ser almacenada en lo que es llamado *Perfiles de Usuario* [17].

El perfil del usuario puede definirse como el conocimiento acerca de la información interesante del usuario. Se distinguen dos tipos de perfiles: los perfiles simples son representados por un grupo de condiciones extraídas de documentos estimados interesantes para ese usuario. Estas condiciones pueden usarse para expandir la consulta original, para recuperar más documentos pertinentes relacionados con lo mismo o temas similares, y los *perfiles extendidos* pueden estar descritos como un tuple de cuatro variables. Una de ellas es la *variable demográfica*, eso está relacionado a los aspectos sociales, como el rango de edad del usuario o el nivel de educación, entre otros. Otra variable es la *identificación*, que está relacionada con la información almacenada en los archivos log del servidor del Web. Luego, las variables *clickstream* que están relacionadas a los pesos de la página, donde si una página tiene un cero el valor, ésta significa que la página no ha sido visitada. La última variable está relacionada a la sesión simple de los perfiles que describimos previamente [17].

Para la obtención de estos perfiles, las técnicas del clustering y reglas de asociación son usualmente aplicadas. A través del proceso de clustering, un grupo de clientes o datos con características similares es inicialmente obtenido sin tener una clasificación previa, se trata de un proceso automático. Los perfiles de usuario derivados de estos grupos pueden ser utilizados para guiar estrategias de marketing según los grupos [21]. Las reglas de asociación descubren relaciones y correlaciones entre artículos en las que la presencia de un artículo o un grupo de

ellas en una transacción implica con un grado de confianza la presencia de otros artículos [4].

El inconveniente principal del manejo de perfiles de usuario en la Web es la falta de conocimiento acerca de la identidad del usuario, idea de la que surgen dos situaciones diferentes. La primera situación tiene que ver con los usuarios sin registrar, casos en el que el perfil de usuario puede identificarse o personalizarse con un grupo social, asignando un perfil general relacionado con unas preferencias mostradas por el usuario mientras navega a través del sitio Web. La segunda situación se refiere a los usuarios registrados: si un usuario es identificado de algún modo, entonces el sitio Web puede estar hecho a la medida según las preferencias del usuario. El sistema sigue la pista al usuario de visitas previas junto con el perfil del usuario, de manera que puede usar esta información para realizar y personalizar el sitio Web.

En ambas situaciones, el previo análisis de los archivos Web log para identificar al usuario es recomendable.

Definición de sesión de usuario

Una sesión de usuario está definida por una secuencia de accesos temporales a un sitio particular de la Web por un usuario [16], [1]. Cada sesión de usuario es una representación lógica de una visita de un usuario a algún sitio Web (usualmente dentro de cierto intervalo de tiempo).

De un archivo log, un conjunto de sesiones de usuarios *S* puede ser definido como *S={s1, s2,...,sm}*, donde cada una de las sesiones si esta definida por un conjunto de páginas visitadas *P={p1, p2,...,pn}*. Por lo tanto, la descripción de una sesión de usuario se puede ver, desde un punto de vista matemático, como una matriz sesión-página SP de dimensiones mxn (Ver ecuación (1)):

$$SP = [w(s_i, p_j)], 1 \leq i \leq m, 1 \leq j \leq n \quad (1)$$

donde w(si; pj) representa el peso de la página p_j en la sesión s_i. Este peso puede tomar valores binarios o puede ser una función de presencia o una función de tiempo de permanencia de la página en la sesión.

Caracterización de sesiones de usuario

De un archivo log inicial, existen varios aspectos para tener en cuenta al momento de identificar cada sesión y predecir a cuál pertenece cada usuario. La etapa de preprocesamiento del archivo log es uno de estos aspectos. El objetivo es limpiar los datos para identificar las sesiones del usuario en una forma más fácil y eficiente. Para limpiar los datos, la información poco relevante es eliminada. Entendemos como la información poco relevante las entradas de gráficos, imágenes, elementos javascript, etcétera, lo que es replicado a través de todo el archivo y generalmente forma parte de los marcos y la composición de páginas dinámicas. Una descripción más detallada de los métodos y técnicas de preprocesamiento pueden ser encontrados en [19].

Otro de los aspectos a considerar es la asociación de una sesión con un cierto usuario. Desde cada web log la entrada del archivo tiene una dirección IP, de modo que la suposición más directa tiene que ver con identificar a cada IP con un usuario diferente. Esto también se puede determinar con una cookie o con una inscripción del usuario en el sitio Web. Sin embargo, estas situaciones no siempre ocurren. La identificación del usuario no es siempre posible, ya que pueden existir casos en los que varios usuarios pueden usar la misma IP a través de un proxy para ganar acceso a un sitio Web. En estas situaciones, podemos identificar una sesión como una secuencia de accesos de la misma IP en un cierto período de tiempo de

retraso grande entre páginas o entre la primera y la última página de la sesión. El método para identificar las sesiones de usuarios es conocido como *timeout* [7] y será explicado con más detalles a continuación.

Del archivo log, una vez que hemos llevado a cabo la etapa de preprocesamiento antes explicado, las sesiones del usuario se organizarán inicialmente asumiendo que cada IP es propia de un único usuario. El siguiente paso es analizar el tiempo de espera para definir las sesiones. Por ejemplo, si una IP en un tiempo A es igual a otra IP en un tiempo B y el tiempo de retraso entre ellos es más pequeño que el intervalo de espera, entonces podemos afirmar que ambas entradas son propias de la misma sesión, de otra manera las IP forman parte de sesiones diferentes. Sin embargo, si una IP en un tiempo A y otra en un tiempo B son diferentes, entonces forman parte de sesiones diferentes, independientemente del tiempo de retraso entre sus accesos.

Formalmente podemos definir un clickstream *R={r1, r2,...,rq}* como una secuencia de click que es registrada en los archivos log cuando el usuario navega por algún sitio Web y realiza clicks sobre imágenes, en algún link, etc. Cada unos de esos click son definidos como r_k, con $1 \leq k \leq q$. Suponiendo que todos los click provienen de una misma IP, podemos caracterizar el clickstream de la siguiente manera:

• *T* es el tiempo de espera calculado como la diferencia entre ambos click en el sitio Web, siendo $T=(t_{k+a} - t_k)$

• Entonces, si $T \leq \beta$, donde β es el máximo tiempo de espera, el click r_k y r_{k+1} es considerado parte de la sesión si. En otro caso, si $T > \beta$, entonces el click r_k es considerado como el final de la sesión s_i, mientras el click r_{k+1} es considerado el comienzo de la sesión $s_{i+1.}$

• Por lo tanto, la duración de la sesión puede ser calculada como:

$$\text{Session time} = \sum_{k=1}^{q} r_k \qquad (2)$$

donde q es el número de clicks en el clickstream de la sesión.

Para tener una más amplia visión de la personalización y la Minería Web, veremos algunos trabajos previos en esta área.

Trabajos Relacionados

En la literatura podemos encontrar algunos enfoques sobre la personalización. Podemos ver una visión general de la personalización en la Web basada en la Minería Web del Uso, y se comentan herramientas y aplicaciones disponibles [12], [22].

También podemos hacer referencia al trabajo de [14] que presenta un sistema de recomendación mediante la integración de técnicas de personalización y de minería, basadas en el comportamiento de usuarios junto a estrategias de marketing aplicado a las tiendas virtuales. El desarrollo de un sistema administrativo de interacción en tiempo real, a través de la interacción de los datos históricos y patrones de uso de accesos en línea de un sitio de comercio electrónico se puede ver en [8].

La más relevante quizás sea [20], donde se presenta una visión general del proceso de personalización basado en la Minería Web del Uso. Se realizan los métodos de procesamiento, integración de datos de fuentes múltiples y las técnicas comunes de descubrimiento de patrones que se aplican a los datos de uso. El objetivo del trabajo es presentar las técnicas de descubrimiento de patrones como clustering, reglas de asociación y descubrimiento de

secuencias de patrones que son usados en el uso de la Web.

Personalización mediante perfiles de usuario con lógica difusa

La lógica difusa parte del principio de que las cosas no son blancas o negras, tal como establece la lógica clásica, sino con tonalidades y con múltiples valores, lo cual se adapta mejor al comportamiento humano.

La lógica difusa se ha convertido en una potente herramienta a la hora de modelar sentencias de lenguaje natural y razonar con las mismas, tal y como lo hace el ser humano, tipo de razonamiento que se conoce como razonamiento aproximado. La teoría de subconjuntos difusos y la lógica difusa, ambos propuestos por L.A. Zadeh [23], constituye los cimientos de la formalización de este tipo de razonamiento.

Los perfiles de usuario son importantes desde el punto de vista de la optimización de Internet y los procesos relacionados con el usuario, como la construcción de consultas, la retroalimentación pertinente del usuario y la construcción flexible de perfiles de los usuarios. La flexibilidad en la construcción de perfiles es necesaria por dos razones: la primera es facilitar el acceso y navegación en la Web, tanto para usuarios expertos como no expertos; y en segundo lugar, para tratar con las ambigüedades de las acciones de los usuarios en la Web.

El conocimiento para la creación de los perfiles lo podemos obtener de la navegación generada por el usuario, la cual está formada por varios componentes correspondientes a las variables conductistas, variables de identificación, variables clickstream y de los documentos estimados interesantes por el usuario.

Estos conceptos son difusos en sí mismos, ya que podemos aproximarnos a la edad, o al nivel de lenguaje o podemos expresar por medio de un cuantificador lingüístico cuan paciente es el usuario. Estas características del usuario pueden ser modeladas a través de diferentes etiquetas lingüísticas, como por ejemplo: {bajo, muy bajo, regular, alto, muy alto} para determinar el nivel de habilidad del usuario para comunicarse en un lenguaje dado.

Los clicks que el usuario hace, nos indican el tiempo que una página Web es visible en el navegador de un usuario, el país de donde está conectado, y nos puede dar una idea acerca del usuario, qué tan paciente es el usuario en la navegación, la edad del usuario, el lenguaje, y así sucesivamente.

Modelo de representación de perfil de usuario

Denotaremos como E el conjunto de los perfiles extendidos. Un perfil extendido e1 pertenece a E, $1 \leq i \leq d$, siendo d el número de perfiles obtenidos, que puede ser representado en la tupla [17]:

$$e_i = (L_i, K_i, z'_i, V_i) \quad (3)$$

donde:

• $L_i = (l_{i1}, l_{i2}, l_{i3}, \ldots, l_{ic})$ es el conjunto de identificación de las variables de los archivos Web log acerca de la identificación del usuario, donde c es el número de variables.

• $K_i = (K_{i1}, K_{i2}, K_{i3}, \ldots, K_{ir})$ es el conjunto de las variables de clickstreams representado por el peso asociado a cada página j disponible, $1 \leq j \leq r$, donde r es el número de variables cliskstreasm consideradas, expresadas en base a lapsos de tiempo en la página; si la página j no es visitada, el valor de k_{ij} es 0.

• $z'_i = \{t'_{i1}, t'_{i2}, t'_{i3}, \ldots, t'_{ia}\}$ donde $z' \in Z$ (Z conjunto de perfiles de usuario), $t'_{ij} \in T$, $1 \leq j \leq a$, donde 'a' es el número de términos en el perfil y que representa una sesión del perfil simple obtenido por la misma sesión.

• $V_i = (v_{i1}, v_{i3}, v_{i3},..., v_{ib})$ representa el conjunto de variables demográficas, siendo 'b' el número de variables a considerar.

Como estas variables pueden ser precisas o imprecisas, las técnicas difusas pueden manejar estos datos con diferentes tipos de granularidades, dependiendo de la variable a modelizar.

Experimentación

El contenido de un perfile de usuario está evidentemente condicionado por los objetivos del mismo dentro de un dominio de aplicación. Es evidente que alguna información contenida en los perfiles puede ser independiente del dominio, como por ejemplo los datos personales.

Proponemos un modelo que nos permitirá la representación automática de perfiles de usuario de algún sitio Web, representado en la figura 1.

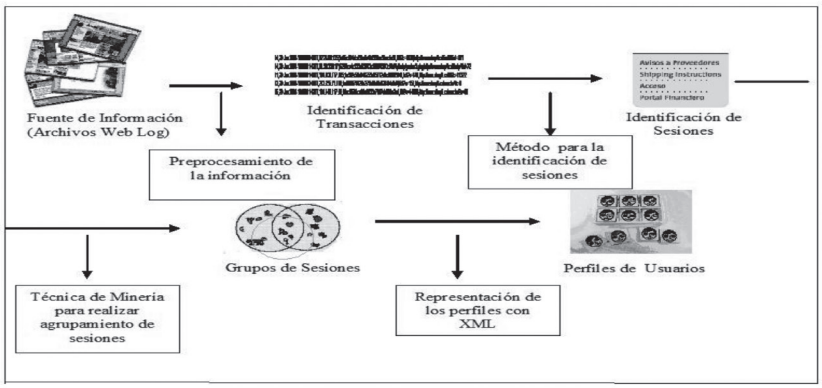

FIGURA 1 MODELO DE REPRESENTACIÓN

Este modelo nos permite analizar información de los servidores Web, siendo esta información obtenida principalmente los ficheros Log que se almacenan en los servidores a medida que los usuarios van navegando por algún sitio Web.

Dentro del análisis propiamente tal del clustering difuso y del tradicional, es importante determinar con anterioridad el número de clusters o grupos implicados en el análisis, claro está que depende mucho del tipo de clustering que se utilice, ya que en algunos no es necesario saber a priori el número de clusters o grupos. Uno de los métodos para determinar el número de clusters o grupos es realizar un análisis a través del clustering jerárquico aplicado al conjunto total de datos o a un conjunto representativo.

A través del análisis del clustering difuso podremos obtener los diferentes conjuntos de las sesiones de usuarios. Cada grupo o cluster posee un centroide que es el elemento que mejor representa al grupo encontrado. Cuando hablamos de un centroide, no queremos decir que se trate de solo "uno", sino que pueden existir varias sesiones que posean las mismas características y representen de mejor manera al grupo o clusters.

De esta manera, al obtener los diferentes centroides de cada grupo encontrado, podremos realizar una representación general de los perfiles de los usuarios. Esta representación reflejará diferentes características relevantes de los usuarios que navegan por el sitio Web, lo que constituye información muy útil al momento de realizar algún tipo personalización dentro del sitio Web.

Para realizar la representación de los perfiles de los usuarios, nos hemos basado en una definición de perfil. Esta definición del perfil de usuario que hemos comentado en la sección 4.1, además, la representaremos a través del lenguaje XML. A continuación veremos la representación general del perfil de usuario a través del lenguaje XML (ver figura 2):

```
<?xml version="1.0" encoding="UTF-8" ?>
<Perfil del Usuario>
      <Identificacion_Usuario>
            <Tipo Tipo= {id_usuario}/>
      </Identificacion_Usuario>
      <Paginas>
            <Pagina Pagina_Visitada="0"> Nombre Página Visitada </Pagina>
            <Pagina Pagina_Visitada="1"> Nombre Página Visitada </Pagina>
                        .
                        .
                        .
            <Pagina Pagina_Visitada="N"> Nombre Página Visitada </Pagina>
      </Paginas>
      <Var_Demograficas>
            <Paciencia>Tiempo de navegación del usuario</Paciencia>
                <Lenguaje> Idioma de la página</Lenguaje>
                <Nivel_Educacional Nivel= nievel de dificultad de la página />
        </Var_Demograficas>
      <Perfil_Simple>
            <Terminos Termino="0">Termino </Terminos>
            <Terminos Termino="1">Termino </Terminos>
                        .
                        .
            <Terminos Termino="N">Termino </Terminos>
      </Perfil_Simple>
</Perfil del Usuario>
```

FIGURA 2 REPRESENTACIÓN GENERAL DE UN PERFIL DE USUARIO EN XML

Comenzaremos explicando la parte del perfil llamada Identificación de Usuario. En esta parte del perfil se realiza la identificación del usuario que se haya conectado al menos una vez al sitio, independientemente de que se haya registrado o no en él.

Existe un problema, que hemos comentado en secciones anteriores, vinculado con el registro de los usuarios. Cuando el usuario está registrado, podemos identificarlo y de esta manera podremos realizar una personalización más adecuada de sus preferencias. En cambio, cuando el usuario no está registrado, se le asigna un perfil general relacionado con sus preferencias mostradas al momento de navegar por la Web.

Dentro de la definición del perfil, también tenemos el campo llamado 'páginas', que se refiere principalmente a las páginas que el usuario ha visitado durante su navegación por el sitio Web. En el campo de la variable demográfica, hemos definido algunas variables que representan algunas cosas relacionadas con perfil del usuario como el lenguaje, el nivel educacional y la paciencia como característica de su comportamiento durante la navegación.

La variable paciencia está relacionada con el tiempo de navegación el usuario por el sitio Web. A estos valores de tiempo les hemos asociado etiquetas lingüísticas para entender de mejor manera el comportamiento de navegación asociado al tiempo. En la figura 3 podemos ver

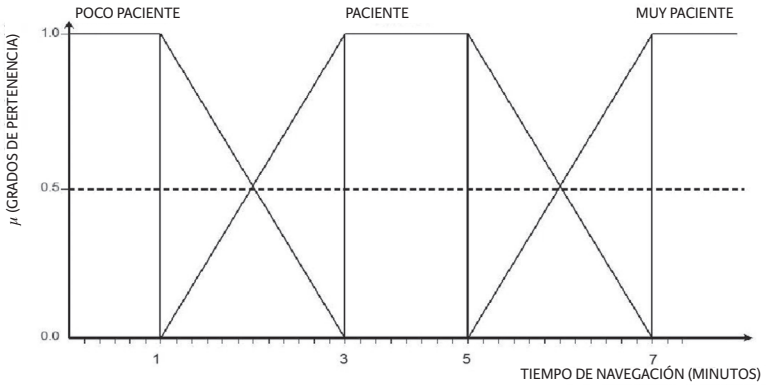

FIGURA 3 ETIQUETAS LINGÜÍSTICAS PARA EL TIEMPO DE NAVEGACIÓN DEL USUARIO

la definición de las etiquetas para la variable de paciencia: Por último, está el campo del perfil simple del usuario, que se relaciona principalmente con las páginas visitadas, la recuperación de información por parte del usuario o, dicho de otra forma, el tipo de información que haya formado parte de su búsqueda. Dentro del código de las páginas Web, existen marcas o tag; entre estas marcas existentes en la páginas podemos resaltar la llamada keyword, que nos indica las palabras claves que se relacionan con el tema o términos importantes dentro del sitio o página Web.

Por ultimo, en la figura 4 podemos ver un ejemplo de resultados obtenidos, en este caso particular, por la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicaciones (ETSIIT) de la Universidad de Granada. El total de perfiles de usuario encontrados fueron de 12. Estos perfiles fueron obtenidos al analizar un conjunto de datos de 15676 transacciones. El objetivo principal del análisis fue identificar si los usuarios que se conectaban eran alumnos o profesores.

De los perfiles obtenidos luego del análisis, podemos decir que la gran mayoría de los usuarios son pacientes o muy pacientes al momento de realizar su navegación por la página de la Escuela. La mayoría corresponden a alumnos y sólo uno de los perfiles corresponde a la de un profesor. Este perfil de profesor se relaciona principalmente con temas como horarios, planes de estudios, convocatorias, entre otros intereses.

Si hacemos referencia a los perfiles de los alumnos, podemos decir que la gran mayoría de los alumnos visitan las páginas de los foros. Dentro de estos foros, podemos encontrar temas relacionados con la programación, la docencia, información general, asignaturas, prácticas, entre otros temas, los cuales se reflejan claramente dentro de los perfiles obtenidos.

De forma general, podemos decir que los términos relacionados con las páginas que visitan los usuarios de la Escuela son principalmente términos relacionados con la Escuela, la Docencia, Temas Generales del Foro, y lo más lógico es que aparezcan términos relacionados con la Ingeniería, la Informática y la Telecomunicaciones.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Perfil5>
  - <Identificacion_Usuario>
      <Tipo Tipo="alumno" />
    </Identificacion_Usuario>
  - <Paginas>
      <Pagina Pagina_Visitada="0">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
      <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
      <Pagina Pagina_Visitada="2">*GET/apps/tablon/HTTP/1.1*</Pagina>
      <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&idhebra=1939HTTP/1.1*</Pagina>
    </Paginas>
  - <Var_Demograficas>
      <Paciencia>Poco Paciente</Paciencia>
      <Lenguaje>Español</Lenguaje>
      <Nivel_Educacional Nivel="" />
    </Var_Demograficas>
  - <Perfil_Simple>
      <Terminos Termino="0">"Index</Terminos>
      <Terminos Termino="1">js/protWindows/themes</Terminos>
      <Terminos Termino="2">tablón</Terminos>
      <Terminos Termino="3">"Ingeniería</Terminos>
      <Terminos Termino="4">Informática</Terminos>
      <Terminos Termino="5">Telecomunicación</Terminos>
      <Terminos Termino="6">Foros</Terminos>
    </Perfil_Simple>
</Perfil5>
```

FIGURA 4
EJEMPLO DE PERFIL OBTENIDO

CONCLUSIONES Y PROYECCIONES

En este artículo hemos revisado los aspectos principales de la Minería Web, centrándonos en el área de la Minería Web del Uso, incluyendo un pequeño estado del arte. También hemos estudiado diferentes aplicaciones de la Lógica Difusa en la Minería Web, la Minería Web del Uso, el Proceso de Personalización y perfiles de usuarios, donde la mayor parte de los elementos son imprecisos por naturaleza.

De esta manera, hemos propuesto un modelo de representación de perfiles de usuario y hemos definido el perfil a través del lenguaje XML. Hemos mostrado un pequeño ejemplo de los resultados obtenidos del análisis de la escuela de Informática y Telecomunicaciones de la Universidad de Granada, siendo la principal característica la utilización del Clustering Difuso que nos permite detectar un comportamiento general en ciertas sesiones y así obtener una mejor agrupación para la obtención de diferentes perfiles de usuarios.

En relación a los trabajos futuros, nuestra intención es profundizar en el área de la personalización, principalmente crear la estructura necesaria para realizar este proceso y estudiar otras técnicas que nos ayuden a realizar la personalización, además de realizar un análisis de correspondencia que podría estar relacionado con los perfiles y las diferentes IP que podemos encontrar entre los datos.

REFERENCIAS

[1] Arlitt, M. Characterizing Web User Sessions, September 1999, Hewlett-Packard Laboratories Technical Report HPL-1999-35R1.

[2] Arotaritei, D., Mitra, S. Web Mining: a survey in the fuzzy framework, 2000, Fuzzy Sets and Systems.

[3] Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms,1981, Plenum, NY.

[4] Carbonell, J., Carven, M., Fienberg, S., Mitchell T., and Yang, Y. Report on the conald workshop on learning from text and the web, 1998, In CONALD Workshop on Learning from Text and The Web, June.

[5] Cooley, R., Mobasher, B., and Srivastava, J. Web mining: Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, 2000, 1-11.

[6] Chakrabati, S. Data Mining for hypertext: A tutorial survey, 2000, ACM SIGKDD Explorations 1(2):1-11.

[7] Chen, M. S., Park, J. S., Yu, P. S. Data mining for path traversal patterns in a web environment, 1996 In Proc. 16th International conference on distributed computing systems, pp. 385-392.

[8] Datta A., Dutta K., VanderMeer D., Ramamritham K. and Navathe S. B. An architecture to support scalable online personalization on the Web The VLDB Journal 10: 104–117 (2001) Springer-Verlag New York 2001.

[9] Delgado, M., Sánchez, D. and Vila, M.A. Fuzzy cardinality based evaluation of quantified sentences, 2000, Int. J. Aprox.Reasoning.

[10] Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M. A. Mining Text Data: Special Features and Patterns. Pattern Detection and Discovery, Proceedings ESF Exploratory Workshop, D.J. Hand, N. Adams and R. Bolton (Eds.), Lecture Notes in Artificial Intelligence Series, 2002, 140-153.

[11] Delgado, M., Marín, N., Sánchez, D. & Vila, M.A., Fuzzy Association Rules: General Model and Applications, 2003, IEEE Transactions on Fuzzy Systems 11, 214-225.

[12] Eirinaki M., and Vazirgianis M., "Web Mining for Web Personalization", ACM Transactions on Internet Tehnology (TOIT), volume 3, issue 1, 2003.

[13] Etzioni, O. The World Wide Web: Quagmire or gold mine, 1996, Comunications of the ACM, 39, 65-68.

[14] García F. J, Gil , A. Personalización y Recomendación en Aplicaciones de Comercio Electrónico. En Avances en Comercio Electrónico, F.García Peñalvo (Ed.), pp.137-148, Editorial Catedral. ISBN: 84-607-5827-3, 2002.

[15] Justicia, C., Martín-Bautista, M. J., Sánchez, D.: Minería de textos: Aplicaciones con lógica difusa, 2004, Actas del Congreso Español de Tecnologías con Lógica Difusa, Jaén, (In Spanish).

[16] Menascé, D., Almeida, V., Fonseca, R. and Mendes, M. A Methodology for Workload Characterization of Ecommerce Sites,1999, Proceedings of ACM Conference on Electronic Commerce (EC-99), Denver, CO, November.

[17] Martín-Bautista M. J., Kraft D. H, Vila M. A., Chen J. and Cruz J.. User profiles and fuzzy logic for Web retrieval issues. Springer Berlin / Heidelberg. ISSN: 1432-7643 (Paper) 1433-7479 (Online) Volume 6, Number 5, pages 365 – 372. August 2002.

[18] Mitra, S., Pal, S.K: Data Mining in Soft Computing Framework: A Survey, 2002 IEEE Transactions on Neural Networks, 3-14.

[19] Mobasher, B., Cooley, R. and Srivastava, J. Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, November 1999, in Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99).

[20] Mobasher, B.: Web Usage Mining and Personalization, 2005, Practical Handbook of Internet Computing M.P.Singh, ed. CRC Press LLC.

[21] Nasraoui, O., Frigui, H., Joshi, A., and Krishnappuram, R. Mining Web access logs using relational competitive fuzzy clustering, 1997, In Proceedings of springs Symposium on Natural Language Proccesing Form the www, Stanford, California. March.

[22] Pierrakos D., Paliouras G., Papalheodorou C. and Spyropoulos C. Web Usage Mining as a tool for personalization:a survey. Kluwer Academic Publishers, vol 13, issue 4, pages 311-372. Hingham, MA, USA, 2003.

[23] Zadeh, L. The concept of linguistic variable and its application to approximate reasoning, 1975, Information Sciences 8, 51-199.